DOCUMENT RESUME

ED 308 245 TM 013 665

AUTHOR van der Linden, Wim J.; Zwarts, Michel A.
TITLE Some Procedures for Computerized Ability Testing. Project Psychometric Aspects of Item Banking No. 13. Research Report 86-10.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE 86
NOTE 48p.; Paper presented in part at the European Conference on Information Technology in Education (Enschede, Netherlands, May 20-23, 1986).
AVAILABLE FROM Mediatheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability Identification; Adaptive Testing; Aptitude Tests; *Computer Assisted Testing; Computer Uses in Education; *Item Banks; *Latent Trait Theory; Test Construction; *Test Items

ABSTRACT

The use of item response theory (IRT) is a prerequisite to successful use of computerized test systems. In item response models, as opposed to classical test theory, the abilities of the examinees and the properties of the items are parameterized separately. Therefore, when measuring the abilities of examinees, the model implicitly corrects for the item properties, and measurement on an item-independent scale is possible. In addition, IRT offers the use of test and item information as local reliability indices defined on the ability scale. The application of IRT in procedures for computerized testing is discussed for: (1) calibrating item banks; (2) automated test construction; (3) adaptive test administration; (4) generating norm distributions; and (5) diagnosing test scores. A 49-item list of references is included. Three figures and one table illustrate the application of IRT to computerized testing. (SDL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED308245

Some procedures for Computerized Ability Testing

Research Report 86-10

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. NELISSEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

W.J. van der Linden
M.A. Zwarts

TM013665

Department of Education

Division of Educational Measurement and Data Analysis

University of Twente

Project Psychometric Aspects of Item Banking No.13

Colophon

Typing : A. Burchartz, L. Padberg
Cover design : M. Driessen, AV-section, University of Twente
Printed by : Central Reproduction Department, University of Twente

Some Procedures for Computerized Ability Testing

Wim J. van der Linden

Michel A. Zwarts

Invited paper to appear in "Technical Advances and Promising Applications of Item Response Theory", Special Issue of the International Journal of Educational Research.

Abstract

For computerized test systems to be operational, the use of item response theory is a prerequisite. As opposed to classical test theory, in item response models the abilities of the examinees and the properties of the items are parameterized separately. Hence, when measuring the abilities of examinees, the model implicitly corrects for the item properties, and measurement on an item-independent scale is possible. In addition, item response theory offers the use of test and item information as local reliability indices defined on the ability scale. In the chapter it is shown how the main features of item response theory have given rise to the development of promising procedures for computerized testing. Among the topics discussed are procedures for item bank calibration, automated test construction, adaptive test administration, generating norm distributions, and diagnosing test scores.

Keywords: Item Response Theory, Computerized Testing, Item Banking

SOME PROCEDURES FOR COMPUTERIZED ABILITY TESTING

For two reasons automation of ability testing has become feasible. First, the large-scale introduction of computers in the schools and the growing interest in their applications in the educational process among teachers and administrators have paved the way for computerized testing. Second, the introduction of item response models in test theory has been a decisive factor. In combination with methods from statistics and operations research already in use for other purposes, these models have features that make them very useful for computerized testing. The integration of computers, item response models, advanced statistics and operations research methods has led to the notion of a computerized test system. The goal of this paper is to focus on these developments.

In order to fully appreciate the distinctive features of item response theory, a comparison with classical test theory is helpful. A main concern of the latter is in the errors of measurement in test scores. Such errors are explained by a linear model decomposing the observed score on a given test into two unknown quantities — the true score and the error of measurement. The history of classical test theory has been one long attempt to draw useful conclusions about these quantities from observed test scores.

An important property of the classical test model is that it is formulated for an observed score on a fixed test. Therefore, conclusions drawn cannot be used in test systems based on item

banking where the option of flexible test construction often leads to tests adapted to individuals, i.e., in general, each individual is administered a unique set of test items.

As opposed to the classical test model, in item response theory the responses of examinees to individual test items are modeled. Also, in these models the ability of the examinee and the properties of the item are parameterized separately. Hence, it is possible to use these models first for the empirical estimation of the item properties from response data of a sample of examinees. In item banking this is called the item calibration stage. Then, once the item parameters are (approximately) known, the parameters for the abilities of the examinees are the only unknown quantities in the model and it can be used as a measurement model: Each time an examinee's ability is of interest, his or her ability parameter value is estimated from the responses to a set of items from the bank. The important point to note is that at the measurement stage the properties of the items administered are represented in the estimation equations for the examinee's ability by parameters with known values (see, for example, Hambleton & Swaminathan, 1985). Therefore, ability measurement is automatically corrected for the properties of the items no matter which set of items from the bank has been administered. For a computerized test system to be flexible this property is a prerequisite.

The remainder of this chapter is organized as follows. First, the notion of a computerized test system is explained. Then some item response models proven to be useful in computerized testing

are considered. The main part of the chapter follows, and consists of an introduction to the application of these models in procedures for computerized testing, namely for calibrating item banks, automated test construction, adaptive test administration, generating norm distributions, and diagnozing test scores. A short discussion of some other procedures in testing suited to automation is included at the end of the chapter.

Computerized Test Systems

A computerized test system is an integrated system for the storage of test items, the construction of tests, and the processing of item responses implemented on a computer. An important feature is the possibility of feedback from the processed item responses to the item bank in the system. Using this facility, the calibrations of the item properties can be updated and, consequently, improvement of the quality of test construction and of adaptive test administration procedures is possible.

The basic elements of a test system are given in the flowchart in Figure 1. The frame shows the activities and data sets in the

---

Insert Figure 1 about here

---

system. In particular, data sets are represented by parallelograms

and activities by circles. Activities consist of operations on data sets. Information flows are indicated by lines which should be read from the top to the bottom. In agreement with this, the data sets at the top of the diagram constitute the input to the system; the sets at the bottom are the output. In the activity Item Bank Processing an item bank is built up from individual test items. Test Construction is an activity in which a test is selected from the item bank such that its specifications in the test order are met as well as possible. Processing of Test Results refers to a transformation of the item responses into interpretable data as requested in the processing order. In Adaptive Testing, items are administered to an examinee one by one, each next item being selected so as to provide the most information at the level of the examinee's ability estimate which is based upon performance on the previously administered items. Flowcharts like the one in Figure 1 play an important role in the application of the ISAC (Information System work and Analysis of Change) methodology to the design of a computerized test system (van Thiel & Zwarts, 1986). For a general description of ISAC, see Lundeberg, Goldkuhl, and Nilsson (1981).

An important data set in Figure 1 is the Item Bank; it can be considered the core of the test system. An item bank is a collection of items structured in two respects: It has a content as well as a psychometric structure. The content structure forms the link between the subject matter and the items. It serves as a classification scheme for the items and is an indispensable device when adding items to or retrieving them from the bank. The

psychometric structure consists of the set of homogeneous ability scales underlying the bank along with the calibrations of the items on these scales. Ideally, the psychometric structure amounts to a partition of the item bank into sets of homogeneous items each nested within a topic from the content structure.

From an information science point of view, a test system is just a data processing system. When designing such a system several things are essential. For instance, it is important to use a systematic design methodology and not an ad hoc approach. Doing so, the risk of design errors is reduced. Also, a standardized approach promotes efficient communication between all parties in the project. The choice of a data base management system is another critical step in the design of a test system. It must be able to cope with the interrelationships within and between various types of data (numerical, text, graphics) in the system and at the same time allow for a fast interaction with its users. The environment in which the test system has to operate determines its design to a considerable extent. It makes a great deal of difference whether a system has to be designed for use by individual teachers (as has been the case, e.g., for PITA; Nitko & Hsu, 1984) or in support of a minimum comptency testing program (see, for example, the Wisconsin Testing Program; Burke, Kaufman, & Webb, 1985). It also makes a great deal of difference whether the system will be used in school districts with a common curriculum or where each school can choose its own curriculum. The best way to adapt a test sytem to its environment is to involve its future users in the development

process, for instance, by following a prototyping strategy. For a more extensive treatment of test systems design, readers are referred to van Thiel and Zwarts (1986).

Some Item Response Models

First, the case of dichotomously scored test items is considered. As the responses of a person a to an item i are considered to be the outcome of a process involving stochastic elements, they can be represented by a random variable U_{ai} taking the value $u_{ai} = 1$ for a correct and $u_{ai} = 0$ for an incorrect response. Let $\theta_a \in R$ denote the ability of person a underlying his/her responses to a homogeneous set of items in the bank. One of the most flexible item response models in use is the 3-parameter logistic model:

$$(1) \qquad P(U_{ai}=1) = c_i + (1-c_i)[1+\exp(-a_i(\theta_a-b_i))]^{-1} ,$$

where $b_i \in R$ and $a_i \geq 0$ are parameters for the difficulty and discriminating power of item i, respectively, and $0 \leq c_i \leq 1$ is a lower asymptote to the probability of a correct response reached when the person guesses blindly (Hambleton & Swaminathan, 1985; Lord, 1980). Most of the procedures for computerized testing in this chapter can be applied for this model. Others, however, capitalize on the properties of a more restrictive model proposed by Rasch (1960):

$$(2) \qquad P(U_{ai}=1) = \exp(\theta_a - b_i)[1+\exp(\theta_a - b_i)]^{-1}.$$

Formally, the model in (2) follows from (1) by imposing the restrictions a_i = const > 0 and c_i = 0. However, statistically the models are quite different due to the fact that (2) belongs to the exponential family of probability distributions (e.g., Andersen, 1980, ch. 6) but (1) lacks this property.

As noted earlier, item response models are appropriate for item banking because, unlike classical test theory, they parameterize items and examinees separately. This is immediately clear from the fact that in (1) and (2) the probability of a correct response, which characterizes an interaction between a person and an item, is decomposed into separate sets of parameters for the person and the item. The Rasch model, however, has an additional property not immediately clear from (2) but rendering it quite appropriate for item banking — the separability of the (maximum-likelihood) parameter estimates. This property says that not only in the model itself but also in its estimation, the parameters can be dealt with independently. More concretely, it implies that the likelihood equations for the estimation of the examinee parameters do not contain any (known or unknown) item parameter, and conversely.

For item banking this is a helpful feature since data from every person, even if he or she has responded to only a small number of items, may be accumulated for updating the item parameter estimates (Choppin, 1968; van der Linden & Eggen, 1986).

A possible criticism of the models in (1) and (2) is their re-

striction to dichotomously scored items. An alternative to (2) still retaining its statistical properties but appropriate for items with more response categories is the Partial Credit Model (Masters, 1982). Let U_{ai} have possible values $u_{ai} = 0, 1, \ldots, m_i$, and let b_{ik} be the difficulty of reaching the k^{th} rather than the $(k-1)^{th}$ response category. The model can then be written as

$$(3) \qquad P(U_{ai}=u_{ai}) = [\exp \sum_{k=0}^{u_{ai}} (\theta_a - b_{ik})][\sum_{\ell=0}^{m_i} \exp \sum_{k=0}^{\ell} (\theta_a - b_{ik})]^{-1}.$$

It should be noted that in this model the rank of the highest response category, m_i, is indexed by i. Therefore the number of categories may vary across the items and the model can be applied to banks of items of mixed format.

The above models are only a small sample of the collection of response models developed over the past three decades. Models have been formulated to handle (1) continuous response formats (Samejima, 1973), (2) linear restriction on the item parameters (Fischer, 1974), (3) multidimensional response data (Fischer, 1974; Reckase & McKinley, 1985), (4) guessing (Verhelst, 1986), and many other types of data. However, this is not the place to explore the wealth of available models more deeply.

Procedures for Computerized Ability Testing

Procedures for computerized ability testing based on item response theory will be considered next. The first procedure deals with an

aspect of item bank calibration, namely optimization of its sampling design. The next two have to do with the use of item response theory in test construction and adaptive testing. How item response theory can be applied in test score interpretation by providing procedures for generating norm distributions and diagnosing response patterns will be discussed in the final section.

Calibrating an Item Bank

At first glance, calibrating an item bank seems to be no more than just estimating the parameters of the items in the bank from a sample of response data. In fact, however, mainly as a result of the following three problems, item bank calibration is not that simple: First, an item bank usually contains more items than an examinee can answer in one administration. Second, it may not be clear beforehand whether the item bank covers one or more ability dimensions. Third, interest is usually not in the abilities of the individual examinees in the sample but in an estimate of the ability distribution in the population of interest (Zwarts, Veldhuizen, & Verhelst, 1986). The first two points will now be discussed in more detail; the third point will be taken up in one of the later sections.

An item bank generally contains hundreds of items, many more than a single examinee can answer. This means that different subsets of items must be administered to different subsets of examinees and that the responses must be analyzed either in one

analysis from an incomplete sample or in separate analyses from complete samples equating the results afterwards. A traditional solution along the former line is multiple-matrix sampling (Shoemaker, 1973), but this procedure is only appropriate for estimating classical test and item parameters. In item response theory the latter approach is usual (see, for example, Hambleton & Swaminathan, 1985; Lord, 1980). Vale (1986) reports results from a study into the accuracy of several designs for linking parameter estimates of multiple sets of items onto a common scale. For the logistic response models estimation of the item parameters from a single sample with an incomplete design is possible, though (Lord, 1974). In addition, for the Rasch model it is exactly known what condition the data from an incomplete sample have to meet to guarantee the existence of unique (maximum-likelihood) estimates (Fischer, 1981). As will be shown now, it is even possible to optimally design calibration samples in the sense of their yielding maximum expected information about the item parameters from the response data.

Figure 2 displays the general shape of a calibration sample. The

---

Insert Figure 2 about here

---

rows in the matrix represent possible subgroups $g = 1,\ldots,G$ in the sample whereas the columns stand for items $i = 1,\ldots,I$ nested in

topics t = 1,...,T. The cells contain the decision variables x_{gi} indicating whether item i is administered to group g (x_{gi} = 1) or not (x_{gi} = 0). Now, optimizing a calibration sample amounts to calculating the values of x_{gi} (g = 1,...,G; i = 1,...,I) such that, subject to possible constraints, some objective with respect to the information about the item parameters is optimized.

In item response theory, a usual measure for the precision of the item parameter estimates is Fisher's information in the sample (e.g., Hambleton & Swaminathan, 1985, chap. 6, or Lord, 1980, chap. 5). To simplify the presentation, it is given here only for the item difficulty parameter σ in the Rasch model; generalizations to other models are straightforward. For an examinee a giving one response to item i Fisher's information measure is defined as

$$(4) \qquad I(b_i;\theta_a) \equiv P(U_{ai}=1)P(U_{ai}=0) \ ,$$

where the probabilities in the right-hand side follow from (2). Now, suppose (empirial or subjective) prior knowledge about the ability distribution in group g is available. Let this knowledge be given by a probability density function $\phi_g(\theta)$. The expected information about σ_i from a random examinee from g is then equal to

$$(5) \qquad E_g\{I(b_i;\theta)\} = \int \phi_g(\theta)I(b_i;\theta)d\theta.$$

Maximization of (5) would be a nice objective for item i, but b_i is unknown and the item bank consists of more than one item.

Therefore, for all items a target distribution for the expected information over the range of possible values of σ is specified. Let $(b_1,\ldots,b_\ell,\ldots,b_L)$ be the possible values in which interest exist. Then a useful target over these points is $(r_1y,\ldots,r_Ly)$, where r_ℓ, $\ell = 1,\ldots,L$, are nonnegative constants of proportionality and y is a dummy variable. An increase in y means that the target distribution is "blown up" but that its form as specified by $(r_1,\ldots,r_L)$ remains intact.

An appropriate optimization model is the following:

(6) maximize y

subject to

$$(7) \qquad \sum_{g=1}^{G} N_g E_\theta(I(b_\ell;\theta)x_{gi} \geq r_\ell y \qquad \ell = 1,\ldots,L$$

$$(8) \qquad \sum_{i=1}^{I} x_{gi} \leq n_g \qquad i = 1,\ldots,I$$

$$(9) \qquad x_{gi} \in \{0,1\} \qquad i = 1,\ldots,I; \quad g = 1,\ldots,G$$

In this model the expected information about all item parameters is maximized subject to the condition that in b_ℓ, $\ell = 1,\ldots,L$, the total expected information from the G groups is not smaller than $r_\ell y$. This is stipulated in (7) where N_g is the number of students drawn at random from group g. The restriction in (8)

guarantees that the test length for group g is not larger than a predetermined constant n_g. The model in (6) - (9) is a linear programming model for which algorithms and computer programs exist to calculate the values of the decision variables x_{gi}, $i = 1,\ldots,I$, $g = 1,\ldots,G$, maximizing the objective function y (e.g., Wagner, 1972). Other models, allowing for a hypothethical ability structure underlying the items, more than one item parameter, or minimization of the sample size, can be found in van der Linden (1986a) and van der Linden and Eggen (1987).

An item bank covering a large number of topics can hardly ever be expected to be unidimensional. Since the dimensionality of a bank is determined by its items as well as the population of examinees it serves, unidimensionality can often be reached by splitting the bank and/or the population. In doing so, the results should be in agreement with the topic structure of the items, but even then a large number of possibilities for finding adequate ability scales will remain. If prior to sampling the response data no hypothesis about the ability structure is available, it has to be found by exploratory methods. However, the price one has to pay is that retrospectively the sample may be less informative about the item parameters than when designed with a hypothesis about the ability structure built into optimization models like the one in (6) - (9).

Test Construction

The traditional way of constructing a test is to write the test items and collect them in a test form. In a computerized test system, a large collection of pretested and calibrated items is available in the item bank and tests construction is reduced to selecting a number of items such that previously established test specifications are met. The purpose of this section is to show that automation of the selection procedure is possible using the concepts of item and test information functions from item response theory.

In classical test theory the reliability coefficient is the index of measurement accuracy for the observed test scores. As a product-moment correlation coefficient, it is population dependent in the sense of representing the accuracy of measurement of the distribution of test scores for the given population of examinees. A change of this distribution will generally lead to a change in the value of its reliability coefficient. Another less favorable aspect of the reliability coefficient is that it does not give any information on the accuracy by which a specific ability level is measured. Both properties of the reliability coefficient are not very practical in item banking where tests often are administered individually or have to be tailored to certain ability levels.

In item response theory the classical concept of reliability is replaced by the concepts of item and test information functions. These can be considered as local measures of reliability indicating the accuracy by which an item or a test measures a specific ability

level. In fact, these information functions are generally known in statistics as Fisher's information in the sample, but are considered as a function of the ability parameter here. Let $I_i(\theta)$ and $I_t(\theta)$ denote the information functions for an item i and test t, respectively. For the Rasch model, $I_i(\theta)$ is defined by (4) as $I_i(\theta) \equiv I(\theta;b_i)$; for other response models comparable definitons hold (see, e.g., Hambleton & Swaminathan, 1985, sect. 6.3). Because of the property of local independence, information functions have the feature of additivity. For a test of n items, it then holds that

$$(10) \qquad \sum_{i=1}^{n} I_i(\theta) = I_t(\theta).$$

The property of additivity immediately suggests the following application in a computerized test system: Together with the items their information functions are stored in the system. If a test has to be selected, the user specifies a target function for the test and the system selects the items from the bank such that the sum of their information functions approximates the target function as close as possible. Requirements with respect to the number of items in the test, their coverage of certain topic areas, and the like, are constaints to be imposed on the selection process.

Theunissen (1985, 1986; see also Theunissen & Verstralen, 1986) was the first to formulate an optimization model to implement this selection process. His objective function was the minimization of the number of items in the test. In the model, decision variables,

x_i, $i = 1,\ldots,I$, are needed to indicate if item i from the bank is included in the test ($x_i = 1$) or not ($x_i = 0$). It follows that the number of items in the test is equal to $\sum_{i=1}^{I} x_i$. A minimal test length certainly is an attractive objective, but without any constaints on the objective function minimization of $\sum_{i=1}^{I} x_i$ would yield $x_i = 0$ for $i = 1,\ldots,I$ as the solution. Hence, additional constraints are needed. Suppose the interest in the test information function can be restricted to its values in the points θ_k, $k = 1,\ldots,K$. Let $I(\theta_k)$ denote the values of the target information function for the test at these points. Then a useful constraint is to require that at each point the value of the sum of the item information functions is as least as large as the value of the target. This leads to the following simple optimization model

$$\text{(11)} \qquad \text{minimize} \sum_{i=1}^{I} x_i$$

subject to

$$\text{(12)} \qquad \sum_{i=1}^{I} I_i(\theta_k)x_i \geq I(\theta_k) , \qquad k = 1,\ldots,K$$

$$\text{(13)} \qquad x_i \in \{0,1\}, \qquad i = 1,\ldots,I$$

This model is again a linear programming model for which algorithms and computer programs exist to solve for the optimal values of x_i, $i=1,\ldots,I$. For efficient heuristic procedures to solve (11) - (13), see Boomsma (1986). The idea of applying linear programming models

in automated test construction has been pursued further in a series of papers. Boekkooi-Timminga (1986a, 1986b) presents optimization models for the simultaneous construction of more than one test, with the construction of parallel tests as a special case. Models with other objective functions than minimization of test length and various practical constraints are given in van der Linden & Boekkooi-Timminga (1987). Some examples of automated test construction can be found in Boekkooi-Timminga & van der Linden (1986).

Adaptive Testing Procedures

At the level of a single examinee, test construction is based on a dilemma. The property of additivity of information functions in (10) shows that a test for a single examinee could be selected optimally from the item bank if his or her ability were known. However, the very reason for testing an examinee is that this quantity is not known. The test construction model in (11) - (13) offers some relief in that it guarantees that wherever the examinee is located on the ability scale, the information from the test does exceed the target values, but the price one has to pay is that for an examinee with a known ability the test generally is longer than necessary. Actually, (11) - (13) is a model for designing group-based tests, not for tests tailored to individual administrations as needed in modern computerized testing applications. A solution to this dilemma is adaptive testing.

In adaptive testing, the decisions about which items to

administer are not made prior to the test administration. Instead, they are made sequentially during the testing session. In doing so, the choice of each next item is based on an ability estimate derived from the responses to the previous items. At first the ability estimates are rough, but at each step further information is gathered allowing the items selected to be more and more on target. Useful references to description of adaptive testing procedures are Hambleton and Swaminathan (1985, sect. 13.3), Lord (1970, 1980, chap. 10), and Weiss (1982). Adaptive testing has been made feasible by the introduction of the computer in testing. Only a computer is able to update the abiltiy estimates and to search the item bank for the next items in such a quick way that the adaptive testing session runs smoothly without disturbing waiting times. The other source of the adaptive testing technology is item response theory. The feasibility of adaptive testing procedures hinges on the availability of larger collections of test items with known properties, procedures for ability estimation, and rules for item selection. How item response theory meets the first two conditons has already been shown in the foregoing; two major rules for item selection in adaptive testing - the maximum-information and a Bayesian sequential rule - will now be discussed.

An obvious rule is the following: The response vector associated with the previous items is used to produce a maximum-likelihood estimate of the examinee's ability. The next item is then selected such that its value for the item information function is maximal at the estimated ability level of the examinee. For the one- and two-

parameter logistic models the item information function is symmetric about $\theta = b_i$ and reaches its maximum at this value. Therefore, the maximum-information rule in this case reduces to selecting the item with b_i at the smallest distance from θ. For the three-parameter model the form of the information function is more complicated and the acutal values of the information functions at θ for all items in the bank have to be calculated to pick out the item with the maximum value.

A disadvantage of the maximum-information rule is that, in particular in the beginning of the procedure, the ability estimates do not need to be finite. This occurs if all item responses are either correct or incorrect and can easily be remedied by administering a few hard or easy items, respectively. An item selection rule without this problem is Owen's (1975) suggestion to apply the Bayesian sequential framework. In this procedure, maximum-likelihood estimation is replaced by the calculation of a posterior distribution for the examinee's ability which serves as the prior distribution for the next step in the procedure. The next item is then found by preposterior analysis: For each item the expected reduction in the variance of the prior distribution is estimated. The item with the largest reduction promises most information about the examinee's ability level and is the optimal choice. With both rules the first step is of critical importance. In the maximum-information approach this is the selection of the first item while in the Bayesian approach it is the first prior. If these are chosen too far from the examinee's actual ability level,

the procedure will be unnecessarily long.

The above two procedures are used most frequently in applications of adaptive testing but in no way exhaust the possibilities. Other options are the up-and-down and Robbins-Monro rules whose applications in adaptive testing have been studied extensively by Lord (1970). In these procedures in each step the ability estimates are changed by a predetermined amount depending only on whether the item is answered correctly or incorrectly. In the up-and-down rule the changes for the correct and incorrect responses are the same in each step whereas in the Robbins-Monro procedure the change decreases as a function of its rank. Adaptive testing procedures have also been studied with a special eye to decision making (Weiss, 1982). A stratified adaptive testing procedure has been proposed by Weiss (1973). The basic idea of this procedure is that dependent on the responses to the test items the examinee may move to different strata of the item bank. A variant of this idea is Lord's (1980) flexilevel test. Adaptive testing implemented at the level of subtests instead of individual items is known as multi-stage testing (Fischer & Pendl, 1980; Lord, 1980).

Generating Norm Distributions

A common practice in ability testing is to provide examinees with information about their relative standing in some norm population. Usual norm populations are examinees of the same age, examinees who are taught the same curriculum, or experts at a given skill. The probability distribution of test scores for a norm population is

known as the norm distribution. The usual way of providing test scores with normative information is to transform them into percentile scores for a relevant norm distribution.

The use of norm distributions has had a long tradition in the practice of standardized testing. In this tradition, abilities have been measured with standardized tests especially developed and pretested for this purpose guaranteeing comparability of test scores by maintaining the same test for a long period for the population of examinees. As a consequence the same norm distribution applies to all examinees, while, in principle, the test score of each examinee can be used for updating the estimates of these distributions. This practice, however, is not feasible for test systems in which item banks are used and tests are tailored to small-scale administration. In such cases it is impossible to estimate norm distributions for all possible tests from the bank, the reason simply being that the number of possible tests is too large. (To illustrate this point: For an item bank consisting of only 33 items, no fewer than 2^{33} different tests are possible just enough to give each inhabitant of the world his/her own test.)

Using item response theory, the problem of establishing norm distributions can be dealt with in a different, and to some extent more elegant fashion. Instead of establishing and updating distributions on the test score variables of all possible tests, this is done for the ability parameter in the model. In this way, a test-independent norm distribution is built up. As soon as a test is selected from the bank, the system is able to generate a norm

distribution using the ability distribution of the relevant population and the values of the item parameters in the test.

It was noted earlier that in item bank calibration the interest usually is not in the individual examinees' abilities but in their distribution in a population. The reason is now clear: Knowledge of this distribution is a prerequisite for generating norm distributions for tests from the bank. In order to be able to estimate an ability distribution, the model has to be extended somewhat. In analysis-of-variance terminology, what is needed is a mixed model for a design with items as a fixed and examinees as a random factor. Item response models can easily be extended into a mixed model by changing over to a model for the marginal probability of a correct response. Let $F_p(\theta)$ denote the distribution function of a correct response as a function of θ for norm population p, wheras $p_i(\theta)$ represents the probability of a correct response as a function of θ as given by the item response model. The marginal probability for item i and population p, Π_{pi}, is given by

$$\Pi_{pi} = \int p_i(\theta) dF_p(\theta) \quad . \tag{14}$$

The ability distribution $F_p(\theta)$ can be approached in different ways. One possibility is to follow a nonparametric approach in which it is approximated by a histogram, the relative frequencies of which are the quantities to be estimated (Bock & Aitkin, 1981; de Leeuw & Vernelst, 1986). Another is to assume a parametric form for the

distribution, for instance, the normal distribution (Andersen & Madsen, 1977; Sanathanan & Blumenthal, 1978; Zwarts & Veldhuizen, 1985) or the more flexible lambda distribution (Verstralen, 1984), and to estimate its parameters. In either case the use of an EM-algorithm may be an appropriate choice for getting maximum-likelihood estimates (Bock & Aitkin, 1981; Zwarts & Veldhuizen, 1985). If students are sampled from different norm populations, more than one distribution has to be estimated. If the distributions differ only in their locations, it is possible to impose a linear model on the population means making parameter estimation more efficient. Alternative models can be compared using a likelihood-ratio test (Zwarts & Veldhuizen, 1985).

Instead of generating a norm distribution for a test from the bank, it is also possible is to work the other way around transforming the ability estimates into percentiles of the distribution on the ability scale. This is computationally less involved but has the disadvantage of reporting normative score interpretations to users of the test system with respect to a distribution they may not understand. Hambleton (1980; see also Hambleton & Swaminathan 1985, sect. 12.4) presents an approach in which this procedure is followed but the ability estimates are used to predict the scores on a standard test from the bank for which normative information is available.

The most surprising thing to note about the procedures in this section is that they can be used to generate normative information for a test even though none of its items was administered to the

sample of examinees from which $F_p(\theta)$ was estimated. Again, this is one of the advantages of using an item response model to calibrate the test items before the item bank is used in the test system.

Diagnosing Test Scores

Providing examinees with information about their relative standing in some norm population usually is not sufficient: Most consumers of test scores want to know what their scores mean. Item response theory also offers a promising application to this problem of test score interpretation.

If a test from the bank has been administered to an examinee, an estimate of his or her value for the ability parameter in the model is available. As the values of the item parameters are also known with sufficient precision, the response model can be used to give estimates of the success probabilities on all items in the bank for the examinee. Thus from a single test score a test system is able to predict the expected performances of an examinee on hundreds of problems typical of the subject area. This offers a very powerful way, known as scale scoring, of providing test scores with a behavioral interpretation. Originally, the idea of scale scoring was presented by Thurstone (1925); however, it was forgotten for over fifty years. Scale scoring was rediscovered for the purpose of reporting results from national assessment studies (Pandey, 1986; Pandey & Carlson, 1983). It has also been used for reporting diagnostic information from achievement tests (Woodcock, 1976). The basic idea of scale scoring is that the ability scale cannot only

be used to calibrate the items but that the locations of items, in turn, also define the ability scale through showing what problems examinees are able to solve with a certain probability. Figure 3 gives a fictitious example of scale score reporting. For a logistic test model without a guessing parameter the examinees are able to

Insert Figure 3 about here

solve all problems below their ability level with a probability larger than 0.50; for problems above their level this probability is smaller. It is an easy job for a computerized test system to print such graphs. In doing so, more precise estimates of success probabilities on the items can be displayed as numerical information in the graph.

Another use of the same principle is to check the response pattern of the items in the test used to obtain ability estimates. For several reasons examinees may not have responded to some of the items as expected. This can be detected by comparing the pattern of actual responses with the expected pattern following from the model. Table 1 gives three obvious examples of suspicious response

Insert Table 1 about here

patterns for the Rasch model. The pattern of examinee A has unexpected responses at the end of the test, both on the easy and the hard items. A possible explanation is that the examinee did not reach the final items and guessed. In pattern B the examinee did not solve the harder items in the beginning of the test; later, however, items even harder were solved. Also, apparently there was no guessing on the harder items in the beginning of the test. This may be an able student who just had a slow start. The examinee of pattern C scored correct responses on some of the harder items but missed some of the easier ones. Since no special pattern can be observed, the best explanation is an examinee of low ability guessing on the hard items. The final pattern is a typical example of an examinee not showing many unexpected responses.

Analyses like those above can be supplied by a test system to its users as an aid in diagnosing test scores. Another application is automated screening of response data in item bank calibration to improve the quality of the item parameter estimates. In all such applications it is assumed that the item response model is the basic model fitting the data, but that for some examinees other factors than the ability of interest may have influenced their responses.

Concluding Remark

The purpose of this chapter to illustrate some promising applications of item response theory in computerized ability testing. Not

all possibilities have been discussed here. Other applications in progress are, for instance, automated item bias detection for subgroups of students (Kelderman, 1986b), interactive setting of performance standards (van der Linden, 1986b), the generation of classical test and item indices for tests from the bank (van der Linden, 1986c), and procedures for automated updating of the item parameter values (Sluyter, 1986). These applications also illustrate now fruitful the use of item response theory in computerized testing can be.

References

Andersen, E.B. (1980). Discrete statistical models with social science applications. Amsterdam: North Holland.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. Psychometrika, 46, 443-459.

Boekkooi-Timminga, E. (1986a). Simultaneous test construction by zero-one programming. Methodika, 1, to appear.

Boekkooi-Timminga, E. (1986b). Algorithms for the construction of parallel tests by zero-one programming (Research Report 86-7). Enschede, The Netherlands: University of Twente, Department of Education.

Boekkooi-Timminga, E., & van der Linden, W.J. (1987). Algorithms for automated test construction. In F.J. Maarse, L.J.M. Mulder, W.P.B. Sjouw, & A.E. Akkerman (Eds.), Computers in psychology: Methods, instrumentation and psychodiagnostics. Lisse: Swets & Zeitlinger.

Boomsma, Y. (1986). Item selection by mathematical programming (Specialistisch Bulletin Nr. 47). Arnhem, The Netherlands: Cito.

Burke, N.W., Kaufman, B.D., & Webb, N.L. (1985). The Wisconsin item bank: Development, operation and related issues. Madison: Wisconsin Department of Public Instruction.

Choppin, B.H.L. (1968). An item bank using sample free calibration. Nature, 219, 870-872.

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. Journal of Educational Statistics, 11, 183-196.

Fischer, G.H. (1974). Einfuehrung in die Theorie psychologischer Tests. [Introduction to the theory of psychological tests]. Bern: Verlag Hans Huber.

Fischer, G.H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. Psychometrika, 46, 59-77.

Fischer, G.H., & Pendl, P. (1980). Individualized testing on the basis of the dichotomous Rasch model. In L.J.Th. van der Kamp, W.F. Langerak, & D.N.M. de Gruijter (Eds.), Psychometrics for educational debates. New York: Wiley, 1980.

Hambleton, R.K. (1980). Latent ability scales, interpretations, and uses. In S. Mayo (Ed.), New directions for testing and measurement: Interpreting test scores (No. 6). San Fransisco: Jossey-Bass.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Nijhoff Publishing.

Kelderman, H. (1986b). Item bias detection using the loglinear Rasch model: Observed and unobserved subgroups (Research Report 86-2). Enschede, The Netherlands: University of Twente, Department of Education.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 39, 219-224.

Lord, F.M. (1970). Some test theory for tailored testing. In W.H. Holtzman (Ed.). Computer-assisted instruction, testing and guidance. New York: Harper & Row.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum.

Lundeberg, M., Goldkuhl, G, & Nilsson, A. (1981). Information systems development: A systematic approach. Englewood Cliffs, N.J.: Prentice Hall.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Nitko, A.J. & Hsu, T.C. (1984). A comprehensive microcomputer classroom testing system. Journal of Educational Measurement, 21, 377-390.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Pandey, T.N. (1986). State of the art of large-scale assessment in the United States. In W.J. van der Linden & J.M. Wijnstra (Eds.), Ontwikkelingen in de methodologie van het onderwijs-onderzoek. Lisse: Swets & Zeitlinger.

Pandey, T.H., & Carlson, D. (1983). Application of item response models to reporting assessment data. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen. Danmarks Peadogogiske Institut.

Reckase, M.D. (1985). Some latent trait theory in a multidimensional latent space. In D.J. Weiss (Ed.), Proceedings of the 1982 item response theory and computerized adaptive testing conference. Minneapolis, MN: University of Minnesota, Department of Psychology.

Samejima, F. (1973). Homogeneous case of the continuous response model. Psychometrika, 38, 203-219.

Sanathanan, L. & Blumenthal, S. (1978). The logistic model and the estimation of latent structure. Journal of the American Statistical Association, 36, 794-799.

Shoemaker, D.M. (1973). Principles and procedures of multiple matrix sampling. Cambridge, Mass.: Ballinger.

Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika,, 50, 411-420.

Theunissen, T.J.J.M. (1986). Optimization algorithms in test design. Applied Psychological Measurement, 10, to appear.

Theunissen, T.J.J.M., & Verstralen, H.H.F.M. (1986). Algorithmen voor het samenstellen van toetsen [Algorithms for constructing tests]. In W.J. van der Linden (Ed.), Moderne methoden voor toetsconstructie en -gebruik. Lisse: Swets & Zeitlinger.

Thurstone, L.L., (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 16, 433-451.

Vale, C.D. (1986). Linking item parameters onto a common scale. Applied Psychological Measurement, 10, to appear.

van der Linden, W.J. (1986a). Een methode voor het optimaliseren van steekproeven in peilingsonderzoek. [A method for optimizing sampling in national assessment studies]. In W.J. van der Linden & J.M. Wijnstra (Eds.), Ontwikkelingen in de metnodologie van het onderwijsonderzoek. Lisse: Swets & Zeitlinger.

van der Linden, W.J. (1986b). A procedure for interactive standard setting. In G.R. Buning, T.J.H.M. Eggen, H. Kelderman, & W.J. van der Linden (Ed.). Het gebruik van net Raschmodel voor een decentraal toetsservicesysteem (Rapport 86-3). Enschede, The Netherlands: University of Twente, Department of Education.

van der Linden, W.J. (1986c). Item banking met een dialoog gebaseerd op klassieke item- en testparameters [Item banking with a dialogue based on classical item and test parameters]. In G.R. Buning, T.J.H.M. Eggen, H. Kelderman, & W.J. van der Linden (Ed.). Het gebruik van net Raschmodel voor een decentraal toetsservicesysteem (Rapport 86-3). Enschede, The Netherlands: University of Twente, Department of Education.

van der Linden, W.J. & Boekkooi-Timminga, E. (1986). Algorithmic test design with practical constraints. Enschede, The Netherlands: University of Twente, Department of Education.

van der Linden, W.J., & Eggen, T.J.H.M. (1986). An empirical Bayes approach to item banking. Applied Psychological Measurement, 10, to appear.

van der Linden, W.J., & Eggen, T.J.H.M. (1987). Algorithms for efficient item bank calibration. In F.J. Maarse, L.J.M. Mulder, W.P.B. Sjouw, & A.E. Akkerman (Eds.), Computers in psychology: Methods, instrumentation and psychodiagnostics. Lisse: Swets & Zeitlinger.

van Thiel, C.C., & Zwarts, M.A. (1986). Development of a testing service system. Applied Psychological Measurement, 10, to appear.

Verhelst, N. (1986). IRT-modellen waar expliciet rekening gehouden wordt met raden [IRT-models explicitly allowing for guessing] (Project TSS, Projectpublicatie nr. 1). Arnhem, The Netherlands: Cito.

Verstralen, H. (1984). Normen bij een Rasch-gecalibreerde item bank [Norms in a Rasch-calibrated item bank]. Tijdschrift voor Onderwijsresearch, 9, 303-316.

Wagner, H.M. (1975). Principles of operations research. London: Prentice-Hall.

Weiss, D.J. (1973). The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Woodcock, R.W. (1976). Key Math: diagnostic arithmetic test. Circle Pines, Minnesota: American Guidance Service.

Zwarts, M.A. & Veldhuizen, N.H. (1985). Grootste-marginale-aannemelijkheidsschatters voor gemengde latente-trekmodellen [Maximum marginal likelihood estimators for mixed latent trait models] (Specialistische bulletin nr. 38). Arnhem, The Netherlands: Cito.

Zwarts, M.A., Veldhuizen, N.H., & Verhelst, N. (1986). Calibreren van item banken in een onvolledig design. [Calibrating item banks with incomplete designs]. In W.J. van der Linden & J.M. Wijnstra (Eds.), Ontwikkelingen in de methodologie van het onderwijsonderzoek. Lisse: Swets & Zeitlinger.

Authors' Note

Portions of this paper were presented at the European Conference on Information Technology in Education, University of Twente, Enschede, The Netherlands, May 20-23, 1986.

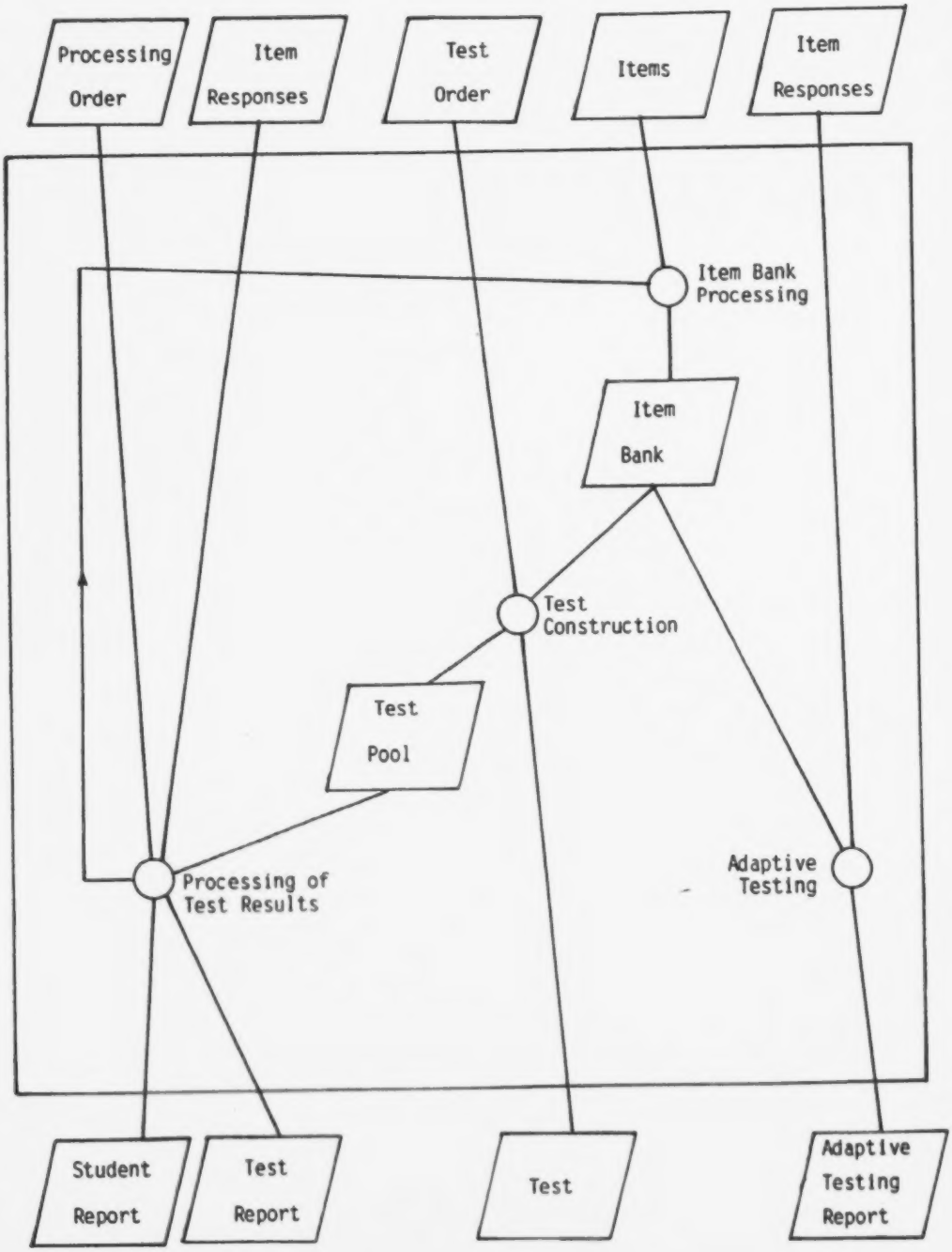
Michel A. Zwarts is at the National Institute of Education (Cito), Arnhem, The Netherlands.

Figure Captions

Figure 1. Flowchart of a computerized test system

Figure 2. The general shape of a calibration sample

Figure 3. A fictitious example of scale score reporting of arithmetic test data


Processing Order
Item Responses
Test Order
Items
Item Responses
Item Bank Processing
Item Bank
Test Construction
Test Pool
Processing of Test Results
Adaptive Testing
Student Report
Test Report
Test
Adaptive Testing Report

Topic		1	. . .	t	. . .	T	
Item	1		. . .	i	. . .		I
Subgroup							
1							
.							
.							
.							
g				x_{gi}			
.							
.							
.							
G							

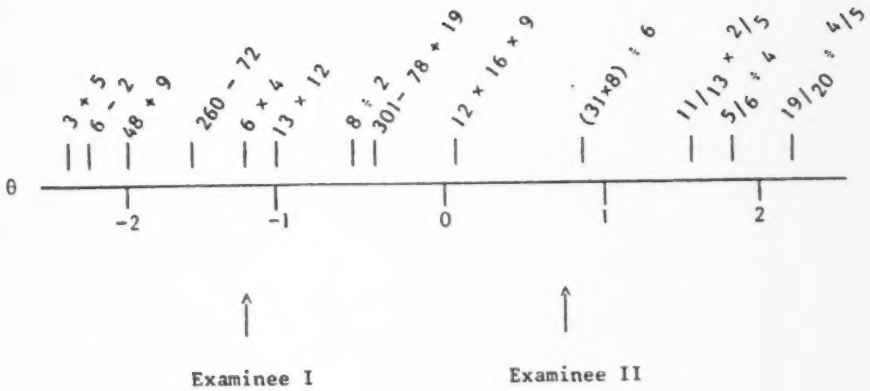

3 × 5
6 − 2
48 ÷ 9
260 − 72
6 × 4
13 × 12
8 ÷ 2
301 − 78 × 19
12 × 16 × 9
(31×8) ÷ 6
11/13 × 2/5
5/6 ÷ 4
19/20 ÷ 4/5
θ
-2
-1
0
1
2
Examinee I
Examinee II

Table 1

Some Response Patterns Suspicious under the Rasch Model

Item	1	2	3	4	5	6	7	8	9	10
b_i	0.75	1.75	-0.50	-0.75	-1.00	-0.25	-2.00	1.50	2.25	2.25
A	0	0	1	1	1	0	0	1	1	0
B	0	0	1	1	1	1	1	1	0	1
C	0	1	1	1	1	0	1	1	0	1
D	1	0	1	1	1	0	1	0	0	0

Titles of Recent Research Reports

RR-86-1 W.J. van der Linden, The use of test scores for classification decisions with threshold utility

RR-86-2 H. Kelderman, Item bias detection using the loglinear Rasch model: Observed and unobserved subgroups

RR-86-3 E. Boekkooi-Timminga, Simultaneous test construction by zero-one programming

RR-86-4 W.J. van der Linden, & E. Boekkooi-Timminga, A zero-one programming approach to Gulliksen's random matched subtests method

RR-86-5 E. van der Burg, J. de Leeuw, & R. Verdegaal, Homogeneity analysis with k sets of variables: An altenating least squares method with optimal scaling features

RR-86-6 W.J. van der Linden, & T.J.H.M. Eggen, An empirical Bayes approach to item banking

RR-86-7 E. Boekkooi-Timminga, Algorithms for the construction of parallel tests by zero-one programming

RR-86-8 T.J.H.M. Eggen, & W.J. van der Linden, The use of models for paired comparisons with ties

RR-86-9 H. Kelderman, Common Item Equating Using the Loglinear Rasch Model

RR-86-10 W.J. van der Linden, & M.A. Zwarts, Some Procedures for Computerized Ability Testing

Research Reports can be obtained at costs from Mediatheek, Faculteit Toegepaste Onderwijskunde, Universiteit Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

A publication by
the Department of Education
of the University of Twente,
P.O. Box 217,
7500 AE Enschede,
The Netherlands

Department of Education